Overview

In order to load data to seqr, you must have a joint called VCF available in your workspace. A joint called VCF is the product of performing joint variant calling on a set of GVCF files. This VCF must be called by either GATK or DRAGEN, and must conform to the requirements specified below. For more information on why seqr requires joint called files, see here.

Generating a joint-called VCF

The GATK team and Terra have created a Terra workspace that allows Terra users to create a joint called VCF for genomes or exomes on GRCh38 starting from unmapped BAMs. Please use this workspace to generate a joint called VCF if you do not already have one.

If you have a joint called VCF that was not produced following GATK's best practices, please confirm the VCF file conforms to the VCF specifications using GATK's ValidateVariants, or VCFTools vcf-validator. The official VCF file specifications can be found here and a general overview of GATK best practices can be found here.

seqr VCF requirements

The following conditions must be satisfied for seqr to import and read a VCF:

1. File meta-information included after the ## string in the header.
    ii. FILTERS that have been applied to the data, for example:

```
Unset

##FILTER=<ID=PASS,Description="All filters passed">
##FILTER=<ID=ExcessHet,Description="Site has excess het value larger
than the threshold">
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=NO_HQ_GENOTYPES,Description="Site has no high quality
variant genotypes">
##FILTER=<ID=high_CALIBRATION_SENSITIVITY_INDEL,Description="Site
failed INDEL model calibration sensitivity cutoff (0.99)">
##FILTER=<ID=high_CALIBRATION_SENSITIVITY_SNP,Description="Site
failed SNP model calibration sensitivity cutoff (0.997)">
```

    iii. Individual format fields, for example:

```
Unset

##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allele Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype Quality">
```

2. 9 columns in the header. The 8 fixed, columns as specified in section 1.3 of the VCF 4.2 Spec plus an additional FORMAT column for individual level fields:

```
Unset

#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
FORMAT
20      1230237 .       A       T       90      PASS    .
GT:AD:GQ
```

3. 3 format fields:
    ii. GT: Genotype, encoded as allele values separated by a slash (ex: "0/0", "./."). seqr does support haploid calls on Y, male non- pseudoautosomal X, or mitochondrion.
    iii. AD: Allele Depth, the number of reads that support each of the reported alleles (ex: "34,0")
    iv. GQ: Genotype Quality, encoded as a phred quality (ex: "99")

seqr VCF Validation

Additionally, seqr performs a series of validations over the VCF to ensure the correctness of several properties:

1. All reference and alternate allele pairs are one of the following Allele Types
    a. Single-nucleotide Polymorphism (SNP)
    b. Multi-nucleotide Polymorphism (MNP)
    c. Insertion
    d. Deletion
    e. Complex Polymorphism
    f. Star Allele (e.g. “alt=*”)
    g. Symbolic Allele (e.g. alt=<INS>)
2. Each variant is only present once, including after multi-allelic variants are split into individual rows.
3. Contigs 1 -> 23 and X are included and have > 100 called variants.
4. The callset contains common coding or non-coding variants that align with the supplied reference genome (GRCh37 or GRCh38) and sequencing type (WES or WGS).
5. The Allele Depth field has length equal to the combined count of REF and ALT alleles on all samples at a site.